CLAIMS

What is claimed is:

1. A system that facilitates enhancement of a speech signal comprising:

an input component that receives a speech signal and pixel-based image data relating to an originator of the speech signal; and,

a speech enhancement component that employs a probabilistic-based model that correlates between the speech signal and the image data so as to facilitate discrimination of noise from the speech signal, the model employing a set of hidden variables representing relevant features, the features being inferred from at least one of the speech signal and pixel-based image data.

2. The system of claim 1, the probabilistic-based model comprising an audio model, the audio model based, at least in part, upon:

$$p(u \mid s) = \prod_k N(u_k \mid 0, \sigma_{sk})$$

$$p(s) = \pi_s$$

$$p(w \mid u) = \prod_k N(w_k \mid hu_k, \phi_k)$$

where u_k is a clean speech signal,

w_k is the speech signal,

s is a state variable of the speech signal, and,

the notation $N(x \mid \mu, \sigma)$ denotes a Gaussian distribution over random variable x with mean μ and inverse covariance σ.

3. The system of claim 1, the probabilistic-based model comprising a video model, the video model based, at least in part, upon:

$$p(l) = const.$$

$$p(v|r) = \prod_i N(v_i | \sum_j A_{ij} r_j + \mu_i, v_i)$$

$$p(y|v,l) = \prod_i N(y_i | v_{i-l}, \lambda)$$

where y is the pixel-based image,

r is a hidden variable,

A is a matrix of weights for the hidden variables r,

l is a location parameter,

v is a hidden clean pixel-based image,

v_{i-l} is shorthand for $v_{\xi}(x_i - x_l)$,

$x(i)$ is the position of the i^{th} pixel,

x_l is the position represented by l, and ,

$\xi(x)$ is the index of v corresponding to 2D position x.

4. The system of claim 1, the probabilistic-based model comprising an audio/video model, the audio/video model based, at least in part, upon:

$$p(r|s) = \prod_j N(r_j | \eta_{sj}, \psi_{sj})$$

where r is a hidden variable,

s is a state variable of the speech signal,

ψ is a precision matrix parameter associated with s, and,

η is a precision matrix parameter associated with s.

5. The system of claim 1, modification of at least one parameter of the probabilistic model being based upon a variational expectation maximization algorithm having an E-step and an M-step.

6. The system of claim 5, the expectation maximization algorithm being based, at least in part, the equation:

$$p(u,s,r,v \mid y,w) \approx q(u \mid s)q(s)q(r \mid s)q(v \mid r,l)q(l)$$

where u is a clean speech signal,
s is a state variable of the speech signal,
r is a hidden variable,
v is a hidden clean pixel-based image,
y is the pixel-based image,
w is the speech signal, and,
l is a location parameter.

7. The system of claim 5, the expectation maximization algorithm being based, at least in part, the equation:

$$h = \frac{\mathrm{Re}\sum_k \phi_k \langle wkEu_k^* \rangle}{\sum_k \phi_k \langle E \mid u_k \mid^2 \rangle}$$

$$\frac{1}{\phi_k} = \left\langle \mid w_k \mid^2 \right\rangle - 2h\,\mathrm{Re}\left\langle w_k E u_k^* \right\rangle + \left\langle E \mid u_k \mid^2 \right\rangle$$

where

$$Eu_k = \sum_s \bar{\pi}_s \bar{\rho}_{sk}$$

$$E \mid u_k \mid^2 = \sum_s \bar{\pi}_s \left(\mid \bar{\rho}_{sk} \mid^2 + \frac{1}{\bar{\sigma}_{sk}} \right)$$

and,
u_k is a clean speech signal,
w_k is the speech signal,
π_s is a prior probability parameter of s,
σ_{sk} is an inverse covariance, and,

8. The system of claim 7, the expectation maximization algorithm being based, at least in part, the equation:

$$
\begin{aligned}
A &= \left\langle Evr^T - EvEr^T \right\rangle \left\langle Err^T - ErEr^T \right\rangle^{-1} \\
\mu &= \left\langle Ev - AEr \right\rangle \\
v^{-1} &= Diag\left\langle Evv^T - AErv^T - \mu Ev^T \right\rangle
\end{aligned}
$$

where "Diag" refers to the diagonal of the matrix, and,

$$
\begin{aligned}
Er &= \sum_s \bar{\pi}_s \bar{\eta}_s \\
Err^T &= \sum_s \bar{\pi}_s \left(\bar{\eta}_s \bar{\eta}_s^T + \bar{\psi}_s^{-1} \right) \\
Ev &= \sum_s \bar{\pi}_s \left(\bar{A} \bar{\eta}_s + \bar{\mu} \right) \\
Evr^T &= \sum_s \bar{\pi}_s \left[\left(\bar{A} \bar{\eta}_s + \bar{\mu} \right) \bar{\eta}_s^T + \bar{A} \bar{\psi}_s^{-1} \right] \\
Evv^T &= \sum_s \bar{\pi}_s \left[\left(\bar{A} \bar{\eta}_s + \bar{\mu} \right) \left(\bar{A} \bar{\eta}_s + \bar{\mu} \right)^T + \bar{A} \bar{\psi}_s^{-1} \bar{A}^T + \bar{v}^{-1} \right]
\end{aligned}
$$

9. The system of claim 8, the expectation maximization algorithm being based, at least in part, the equation:

$$
\begin{aligned}
\eta_{sj} &= \left\langle \bar{\eta}_{sj} \right\rangle \\
\frac{1}{\psi_{sj}} &= \left\langle \left(\bar{\eta}_{sj} - \eta_{sj} \right)^2 + \left(\psi_s^{-1} \right)_{jj} \right\rangle
\end{aligned}
$$

10. The system of claim 1, the image data comprising information associated with an appearance of the lips of the originator of the speech signal.

11. The system of claim 1, wherein the speech component tracks the lips of the originator of the speech signal in order to facilitate discrimination of noise from the speech signal.

12. The system of claim 1, the input component further comprising a frequency transformation component that receives windowed signal inputs, computes a frequency transform of the windowed signals, and provides outputs of frequency transformed windowed signals to the speech enhancement component.

13. The system of claim 12, further comprising a windowing component that applies an N-point window to the speech signal and provides the windowed signal inputs to the frequency transformation component.

14. The system of claim 1, further comprising at least two audio input devices that provide speech signals.

15. The system of claim 1, the probabilistic-based model being trained, at least in part, during operation of the system.

16. The system of claim 1, the features comprising at least one of a speech state and lip motion.

17. The system of claim 1, wherein the model incorporates an additional degree of freedom that models image translation.

18. A method facilitating enhancement of a speech signal comprising:
receiving a speech signal;

receiving a pixel-based image data relating to an originator of the speech signal; and,

generating an enhanced speech signal based, at least in part, upon a probabilistic-based model that correlates between the speech signal and the image data so as to facilitate discrimination of noise from the speech signal.

19. The method of claim 18 further comprising providing an output associated with the enhanced speech signal.

20. A data packet transmitted between two or more computer components that facilitates enhancement of a speech signal, the data packet comprising:

an enhanced speech signal, the enhanced speech signal being based, at least in part, upon a probabilistic-based model that correlates between speech signal and image data related to an originator of the speech signal so as to facilitate discrimination of noise from the speech signal.

21. A computer readable medium storing computer executable components of a system that facilitates enhancement of a speech signal comprising, comprising:

an input component that receives a speech signal and pixel-based image data relating to an originator of the speech signal; and,

an speech enhancement component that employs a probabilistic-based model that correlates between the speech signal and the image data so as to facilitate discrimination of noise from the speech signal.

22. A system that facilitates enhancement of a speech signal comprising:

means for receiving a speech signal and pixel-based image data relating to an originator of the speech signal; and,

means for enhancing the speech signal, the means for enhancing employing a probabilistic-based model that correlates between the speech signal and the image data so as to facilitate discrimination of noise from the speech signal.